Extraction de sous-parties ciblées d'une ontologie généraliste pour enrichir une ontologie particulière

Fayçal Hamdi, Brigitte Safar, Chantal Reynaud

LRI, Bât. 650, Université Paris-Sud 11
91405 Orsay Cedex France
{faycal.hamdi, safar, chantal.reynaud}@lri.fr,
http://www.lri.fr/~hamdi

Résumé. Différentes ressources ontologiques généralistes de très grande taille ont été développées de façon collective et sont aujourd'hui disponibles sur le web. Ainsi l'ontologie YAGO est une énorme base de connaissances décrivant plus de 2 millions d'entités. Afin de tirer parti de ce gigantesque travail collectif, nous montrons comment en extraire des sous-parties thématiquement focalisées pour enrichir une autre ontologie, dite cible, de taille plus limitée mais de domaine centré sur une application particulière [1].

1 Introduction

L'ontologie YAGO (Yet Another Great Ontology) est une énorme base de connaissances, développée à l'Institut d'informatique Max Planck de Sarrebruck, décrivant plus de 2 millions d'entités, des personnes, des organisations, des villes et contenant plus de 20 millions de faits sur ces entités. Ses données proviennent de Wikipedia et sont structurées à l'aide de WordNet. De par sa taille et ses multi-thématiques, YAGO est difficile à utiliser directement pour des applications particulières. Afin de tirer parti plus aisément de ce gigantesque travail collectif, l'objectif de cet article est de montrer comment on peut en extraire des sous-parties thématiquement focalisées pour enrichir une autre ontologie dite cible, de taille plus limitée mais de domaine centré pour une application particulière.

Pour procéder à l'extraction des sous-parties pertinentes de YAGO, nous adaptons l'outil TaxoPart (Hamdi et al., 2009b), développé pour partitionner des ontologies de grande taille avant de les aligner. L'enrichissement de l'ontologie cible est ensuite effectué en procédant par alignement des sous-parties extraites de YAGO avec l'ontologie à enrichir, en utilisant l'outil d'alignement TaxoMap (Hamdi et al., 2009a). Les concepts extraits de YAGO apparaissant dans l'alignement produit sont considérés comme des concepts candidats pour l'enrichissement, liables à ceux de la cible par les relations établies dans les mappings.

Nous commençons par rappeler en section 2, le fonctionnement de TaxoPart, avant de présenter, en section 3, les adaptations faites pour l'utiliser au mieux dans ce contexte. Les résultats d'une expérimentation menée sur une ontologie topographique sont présentés en section 4, les travaux proches et nos conclusions en section 5.

1. Ce travail est financé par l'Agence Nationale de la Recherche (ANR) au travers du projet ANR-07-MDCO-005 pour la création, la comparaison et l'exploitation d'ontologies géographiques hétérogènes (http://geonto.lri.fr/).

2 L'outil de partitionnement TaxoPart

TaxoPart a été développé pour partitionner conjointement deux ontologies de grande taille en blocs de taille plus acceptable pour les outils d'alignement d'ontologies. TaxoPart utilise un algorithme de classification hiérarchique agglomératif, l'algorithme PBM (Hu et al., 2006) qui regroupe itérativement dans un même bloc, les concepts jugés proches suivant une mesure de similarité ne s'appuyant que sur la position relative des concepts au sein de leur ontologie. PBM s'arrête quand tous les blocs construits ont atteint la taille limite, fixée au départ, à partir de laquelle les blocs ne peuvent plus être regroupés.

Pour prendre en compte l'objectif d'alignement, TaxoPart commence d'abord par partitionner l'ontologie la plus structurée, dite cible, ce qui permet d'obtenir un ensemble de blocs sémantiquement cohérents si cette ontologie est effectivement bien structurée.

Dans une deuxième étape, TaxoPart force la décomposition de la deuxième ontologie dite source, à suivre la décomposition de la première. Pour ce faire, l'ensemble des concepts des deux ontologies qui ont les mêmes labels, et que nous appellerons les *ancres*, sont tout d'abord identifiés, avec une mesure lexicale simple et peu coûteuse. Pour chaque bloc B_{Ci} de la cible, les concepts de la source qui correspondent aux ancres appartenant à ce bloc sont regroupés en un ensemble qui formera le noyau d'un futur bloc de la source NB_{Si} (cf. Fig 1). L'introduction de ces noyaux dans le processus conduit PBM à forcer le regroupement des concepts de la source structurellement proches d'une des ancres du noyau à rejoindre le même bloc.

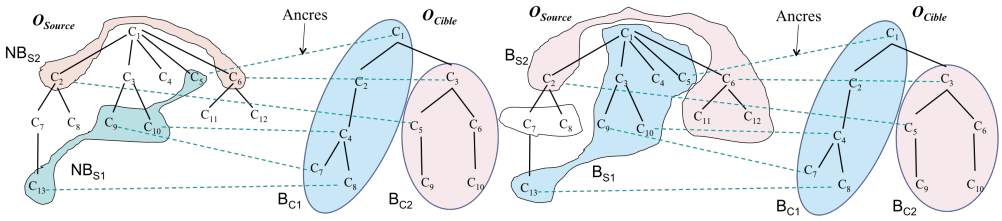


FIG. 1 – *Les noyaux NB_{Si} associés aux B_{Ci}*

FIG. 2 – *Partition de O_S autour des NB_{Si}*

Cette méthode permet de décomposer deux ontologies en deux ensembles de blocs dont certains peuvent naturellement être regroupés par paire (B_{Ci}, B_{Si}) telles que les blocs de chaque paire sont constitués d'un ensemble d'ancres et d'autres concepts a priori proches, au moins structurellement dans leur ontologie respective, des ancres considérées. Les blocs d'une même paire regroupent donc a priori les concepts des deux ontologies concernant les mêmes sous-parties thématiques et peuvent donc être alignés deux à deux (cf. Fig. 2).

Etant donné un certain nombre de blocs cibles, TaxoPart permet, par construction, de centraliser dans le même nombre de blocs sources, les concepts de la source proches de ceux présents dans chacun des blocs cibles. Dans un contexte où l'ontologie source est beaucoup plus volumineuse que l'ontologie cible, si la taille des blocs sources à construire est limitée, les concepts de la source éloignés des ancres vont s'agglomérer dans des blocs indépendants. Ce mécanisme permet donc de n'extraire de l'ontologie source, ici la volumineuse source YAGO, que les sous-parties dont les thématiques sont proches de celles exprimées dans les différents blocs de la cible et de laisser le reste de l'ontologie se décomposer dans des blocs qui ne seront pas examinés.

Pour utiliser au mieux cet outil avec un objectif d'extraction, deux adaptations sont nécessaires. La première consiste à s'assurer de la bonne décomposition de l'ontologie cible puisqu'elle va servir de guide à l'extraction. Nous avons considéré qu'elle pouvait être décomposée manuellement en blocs de tailles variées mais de thématiques ciblées. Cette démarche manuelle se justifie d'une part par le fait que l'ontologie cible est supposée être de taille réduite, et d'autre part, parce qu'elle est a priori bien connue de ses concepteurs et que cette décomposition sera éventuellement réutilisée, i.e. chaque fois qu'on voudra enrichir l'ontologie avec les concepts d'une nouvelle source potentielle. La deuxième adaptation concerne la phase de calcul des ancres. Dans une ontologie de domaine ciblée, il est très rare que deux concepts différents portent un même label. En revanche, dans une ontologie généraliste comme YAGO, i.e. qui couvre plusieurs domaines à la fois, ce phénomène est fréquent. Par exemple, pour le concept `col` de l'ontologie cible, dont le label est traduit en anglais par `pass`, on trouve dans YAGO, 11 concepts distincts portant ce même label. Il y a donc une ambiguité qu'il faut résoudre et qui nécessite d'adapter la phase de calcul des ancres.

3 Algorithme de calcul des ancres

Nous appellerons **ancre** un concept d'une ontologie mis en correspondance avec un unique concept d'une autre ontologie, de même label et de même sens que lui. Nous parlerons d'une paire d'ancres, quand nous ferons référence aux deux concepts mis en correspondance. Cette définition fait apparaître 3 caractéristiques pour la paire de concepts considérés : ils ont le même label, des sens jugés proches et leur mise en correspondance est unique. Avant d'identifier effectivement une paire d'ancres, il faut vérifier successivement ces 3 caractéristiques. Nous parlerons donc de *paire d'ancres potentielle*, pour une paire de concepts ayant le même label, puis de *paire d'ancres plausible*, si les sens des concepts d'une paire potentielle ont été considérés comme proches. Un concept cible donné pouvant appartenir au départ à plusieures paires d'ancres *potentielles* puis *plausibles*, nous qualifierons de *valide*, l'unique paire d'ancres *plausible* retenue par notre algorithme.

Dans un premier temps, on construit l'ensemble des paires d'*ancres potentielles* où chaque paire associe un concept de la cible au(x) concept(s) de YAGO de même label que lui. On évalue ensuite la proximité des sens des concepts d'une paire potentielle en comparant les voisinages de chacun des concepts dans son ontologie respective. Dans notre contexte, le voisinage d'un concept contiendra ses pères et grand pères, ses fils et petits-fils, et ses frères. Un concept de la source ne pourra être reconnu comme une *ancre plausible* d'un concept de la cible que si son voisinage partage au moins une autre ancre avec le voisinage du concept cible.

Si une ancre de la cible n'a aucune ou au contraire, plusieurs ancres plausibles, on effectuera une deuxième passe, dans laquelle on essayera de tirer profit de l'information plus large donnée par les blocs. Au lieu de comparer le voisinage d'une ancre potentielle de la source avec le seul voisinage du concept cible correspondant, on le comparera à l'ensemble des concepts cibles du même bloc pour lesquels on a trouvé une ancre valide dans la première passe. L'idée est que si les deux ontologies ne sont pas structurées selon le même point de vue mais que les blocs de la cible ont une thématique suffisamment focalisée, le fait de retrouver des concepts cibles dans le voisinage d'une *ancre potentielle* de la source, peut suffire à lever l'ambiguité sur cette ancre, même si les concepts trouvés ne sont pas présents dans le voisinage immédiat

de l'ancre cible à laquelle on le compare. Le nombre de concepts cibles trouvés dans le voisinage d'un concept source permettra de choisir entre deux ancres plausibles qui partageaient le même nombre de concepts proches ou permettra de valider une ancre qui ne l'était pas dans la passe précédente.

Une fois identifiées les ancres valides des différents blocs de la source, le partitionnement de l'ontologie s'effectue comme décrit précédemment. La dernière étape consiste à aligner les blocs d'une même paire (B_{Ci}, B_{Si}) et à utiliser les alignements pour l'enrichissement.

4 Expérimentations

Nous avons testé l'algorithme présenté, en utilisant comme cible, une ontologie topographique de taille réduite (600 concepts), que nous avons décomposée manuellement en 10 blocs de tailles variées mais de thématiques ciblées : entités topographiques naturelles, routes, voies ferrée, habitations, éléments du patrimoine, lieux ou bâtiments dédiés au tourisme, etc.

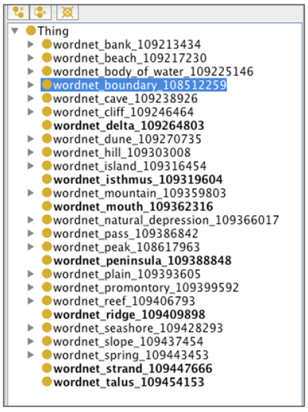
N°du bloc	0	1	2	3	4	5	6	7	8	9
Concepts cibles	102	62	60	24	68	23	69	110	69	30
Ancres cibles potentielles	51	33	38	11	37	13	39	90	43	24
Ancres potentielles de YAGO	120	97	171	44	117	27	142	371	107	73
Ancres valides	22	11	23	9	21	4	18	60	24	15

TAB. 1 – *Tests sur l'ontologie topographique*

On voit par exemple dans le tableau 1 ci-dessus que sur les 110 concepts du bloc cible n°7, qui correspond au bloc regroupant les "entités topographiques naturelles", 90 concepts participent à au moins une des 371 paires d'ancres potentielles pouvant être construites avec les concepts de YAGO. Sur ces 90 *ancres cibles potentielles*, seules 60 ont finalement été retenues comme des *ancres valides*, et nous avons vérifié que leur correspondant dans YAGO relevaient tous du domaine topographique, à une ou deux exceptions près. Nous présentons fig. 3 le bloc construit autour de ces 60 ancres en fixant à 200 concepts la taille limite maximale de fusion. La visualisation effectuée avec le logiciel *Protégé* présente les concepts directement attachés à la racine du bloc, les nœuds ayant des descendants sont précédés d'un triangle, et les nœuds feuilles sont en gras.

Cette visualisation permet à l'expert de vérifier le contenu du bloc avant de l'utiliser pour l'enrichissement.et de supprimer si nécessaire, une éventuelle branche hors domaine introduite dans le bloc à partir d'une fausse ancre, comme ici la branche `boundary`, surlignée en bleu dans la figure. Dans l'exemple, une fois otée la branche hors domaine, le bloc contient 167 concepts, donc 107 concepts différents des ancres mais appartenant à la même thématique.

L'enrichissement s'effectue en utilisant TaxoMap pour aligner chaque bloc source extrait de Yago, avec le bloc correspondant de la cible. TaxoMap met en œuvre un ensemble de techniques terminologiques et structurelles pour générer des mappings exprimés par des relations d'équivalence (*isEq*), de subsomption (*isA* ou *isMoreGnl*) ou de proximité (*isClose*). Par exemple, une des techniques s'appuie sur les mappings d'équivalence identifiés entre un



bayou *isA* lake
lagoon *isA* lake
loch *isA* lake
lough *isA* lake
oxbow lake *isA* lake
reservoir *isA* lake
tarn *isA* lake

flat *isA* plain
llano *isA* plain
moor *isA* plain
peneplain *isA* plain
snowfield *isA* plain
steppe *isA* plain
tundra *isA* plain

branch *isA* stream
headstream *isA* stream
rivulet *isA* stream
tidal river *isA* stream

FIG. 4 – *Mappings utilisables pour l'enrichissement*

FIG. 3 – *Bloc n°7*

concept source et un concept cible, pour générer des mapping de subsomption entre les spécialisations du concept source et le concept cible. Ainsi, à partir des 60 paires d'ancres du bloc n°7, 75 mappings de subsomption seront générés avec cette technique. Quelques exemples sont présentés figure 4. Après validation, l'expert pourra s'appuyer sur les relations établies dans les mappings, pour introduire directement dans la cible certains de ces 75 concepts.

5 Travaux proches et conclusion

Différents travaux portent sur le partitionnement d'ontologies de grande taille. Ainsi Stuckenschmidt et Schlicht (2009) ou Grau et al. (2006) décomposent une ontologie en sous-blocs (ou *îlots*) indépendants les uns des autres, pour faciliter en toute généralité différents traitements (maintenance, visualisation, validation ou raisonnement). Ces décompositions satisfont des critères de taille, d'indépendance, ou de complétude des raisonnements au sein des blocs alors que nous souhaitons ici être guidés par le contenu.

Les travaux les plus proches du nôtre pour cet aspect sont ceux de Noy et Musen (2009) et de Seidenberg (2009), qui permettent à l'utilisateur d'extraire des portions d'ontologies (ou *Vues*) centrées autour d'un ou de plusieurs concepts, en spécifiant les relations entre concepts intéressantes à extraire et la profondeur des sous-arbres extraits. L'accent est ici mis sur les problèmes de cohérence des segments extraits suivant les constructeurs OWL utilisés.

Notre objectif est différent. Nous ne nous appuyons pour l'extraction que sur les relations de subsomption. Nous ne cherchons à extraire que des concepts nouveaux mais relevant d'une thématique précise, définie par un ensemble de concepts, et nous rencontrons de ce fait des problèmes de gestion d'ambiguité non abordés dans les travaux précédents.

Pour traiter les problèmes de polysémie, Zablith et al. (2010), qui se placent aussi dans un contexte d'enrichissement d'ontologies, travaillent concept par concept. Ils évaluent la pertinence de l'introduction d'une nouvelle relation < c_s $relation$ c_c> en comparant le contexte du concept c_s dans son ontologie d'origine avec celui de l'ontologie cible O_C de façon à identifier les concepts partagés, i.e. de mêmes labels. Si des concepts partagés existent et qu'ils vérifient certaines relations structurelles avec les concepts intervenant dans la nouvelle relation à intro-

duire, celle-ci est jugée pertinente pour l'enrichissement. Gracia et al. (2007) ont une approche similaire mais valident les concepts partagés en vérifiant l'existence dans WordNet d'un sysnset partagé par les généralisants de c_s et de c_c dans leur ontologie respective.

En conclusion, nous remarquons qu'en exploitant la présence conjointe de plusieurs termes pour préciser leur sens, au lieu de travailler concept par concept, notre approche permet d'effectuer une désambiguisation collective comme proposée par Cucerzan (2007). Ceci nous permet d'aider un expert à extraire d'une grande ressource ontologique des sous-parties thématiquement ciblées pour enrichir son ontologie particulière, Les résultats obtenus dans les expérimentations sont encourageants.

Références

Cucerzan, S. (2007). Large-scale named entity disambiguation based on wikipedia data. In *Empirical Methods in Natural Language Processing*, pp. 708–716.

Gracia, J., V. Lopez, M. D'Aquin, M. Sabou, E. Motta, et E. Mena (2007). Solving semantic ambiguity to improve semantic web based ontology matching. In *Ontology Matching*.

Grau, B. C., B. Parsia, E. Sirin, et A. Kalyanpur (2006). Modularity and web ontologies. In *Proceedings of KR2006, Lake District, UK, June 2–5, 2006*, pp. 198–209.

Hamdi, F., B. Safar, N. B. Niraula, et C. Reynaud (2009a). Taxomap in the oaei 2009 alignment contest. In *Ontology Matching (OM-2009), Chantilly, USA, October 25, 2009*.

Hamdi, F., B. Safar, H. Zargayouna, et C. Reynaud (2009b). Partitionnement d'ontologies pour le passage à l'échelle des techniques d'alignement. In *EGC*, pp. 409–420.

Hu, W., Y. Zhao, et Y. Qu (2006). Partition-based block matching of large class hierarchies. In *Asian Semantic Web Conference- ASWC*, pp. 72–83.

Noy, N. F. et M. A. Musen (2009). Structure-based partitioning of large ontologies. In *Modular Ontologies, LNCS 5445*, pp. 245–260.

Seidenberg, J. (2009). Web ontology segmentation : Extraction, transformation, evaluation. In *Modular Ontologies, LNCS 5445*, pp. 211–243.

Stuckenschmidt, H. et A. Schlicht (2009). Traversing ontologies to extract views. In *Modular Ontologies, LNCS 5445*, pp. 187–210.

Zablith, F., M. D'Aquin, M. Sabou, et E. Motta (2010). Using ontological contexts to assess the relevance of statements in ontology evolution. In *EKAW 2010*.

Summary

Various large and general ontological resources have been built in a collaborative way and are today publicly available on the web. YAGO is an illustration of such resource. It is a huge semantic knowledge base which has more than 2 millions entities. Our aim is to take advantage of all this work. We show how to extract sub-parts of YAGO focusing on particular topics and how to reuse them in order to enrich a target ontology of a more limited size but centered on a specific application domain.